TOREE: Evaluating Topic Relevance of Student Essays for Chinese Primary and Middle School Education

Xinlin Zhuang[1], Hongyi Wu[1], Xinshu Shen[1], Peimin Yu[3], Gaowei Yi[4], Xinhao Chen[1]
Tu Hu[1], Yang Chen[1], Yupei Ren[1,2], Yadong Zhang[1], Youqi Song[1], Binxuan Liu[1], Man Lan[1,2]*

[1] School of Computer Science and Technology, East China Normal University, Shanghai, China
[2] Shanghai Institute of AI for Education, East China Normal University, Shanghai, China
[3] School of International Chinese Studies, East China Normal University, Shanghai, China
[4] College of Engineering, Ocean University of China, Shandong, China
xinlinzhuang@stu.ecnu.edu.cn mlan@cs.ecnu.edu.cn

Abstract

Topic relevance of an essay demands that the composition adheres to a clear theme and aligns well with the essay prompt requirements, a critical aspect of essay quality evaluation. However, existing research of Automatic Essay Scoring (AES) for Chinese essays has overlooked topic relevance and lacks detailed feedback, while Automatic Essay Comment Generation (AECG) faces much complexity and difficulty. Additionally, current Large Language Models, including GPT-4, often make incorrect judgments and provide overly impractical feedback when evaluating topic relevance. This paper introduces **TOREE** (**To**pic **Re**levance **E**valuation), a comprehensive dataset developed to assess topic relevance in Chinese primary and middle school students' essays, which is beneficial for AES, AECG and other applications. Moreover, our proposed two-step method utilizes TOREE through a combination of Supervised Fine-tuning and Preference Learning. Experimental results demonstrate that TOREE is of high quality, and our method significantly enhances models' performance on two designed tasks for topic relevance evaluation, improving both automatic and human evaluations across four diverse LLMs. [1]

1 Introduction

Topic relevance of an essay requires that the composition not only revolves around a clear theme but also its content aligns with the requirements of essay prompt, which is a critical aspect in evaluating the quality of an essay, as shown in Figure 1. Automating the assessment of topic relevance not only effectively reduces teachers' workload but also efficiently enhances students' writing skills, which holds significant educational implications.

---

*Corresponding author.
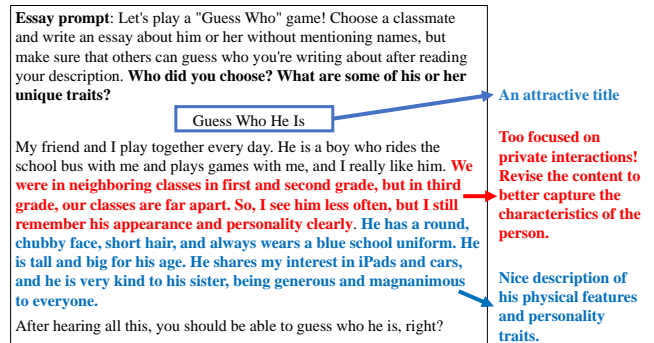[1]Our code and dataset are released at https://github.com/cubenlp/TOREE.

Figure 1: An explanation of topic relevance for a grade 3 essay. Texts highlighted in **red** indicate weakness and provide detailed suggestions, whereas texts highlighted in **blue** point out parts with accurate topic relevance.

Current research on automatic essay evaluation predominantly focuses on the development of Automatic Essay Scoring (AES), which aims to determine holistic and trait-specific rankings, or discrete scores to essays by analyzing a range of textual features, such as coherence, grammatical correctness, and expressive clarity (Jiang et al., 2023; Song et al., 2020; Farag et al., 2018). However, existing research in this field, while considering *topic* as an aspect of evaluation, has neglected the importance of topic relevance (He et al., 2022) (See Appendix A for detailed explanations of these two aspects in Chinese essay evaluation). Furthermore, despite the efficiency in scoring offered by AES systems, it is the detailed comment and feedback that is particularly beneficial for students to directly refine their essays and strengthen their writing abilities.

Automatic Essay Comment Generation (AECG) (Liu et al., 2023) seeks to provide students with constructive feedback tailored to improve particular areas of their writing as delineated by the strengths and weaknesses highlighted in the comments, which is commonly adopted by teachers (Liu et al., 2017). Nevertheless, the intersection of complexity and difficulty in AECG results in a

EduChat (a)	这篇作文的中心与切题方面表现**良好**。作者**选取了一个具体的人物——尤唯**，通过描写她的品质和行为，深刻地展现了她的精神风貌。尤唯是一个**品学兼优、尊敬师长、文静聪慧**的女孩，她的友谊和帮助让作者感动。作者通过描写自己与尤唯的互帮互助，表达了**对友谊的珍视和对未来的期许**。整篇作文紧扣主题，情感真挚，给人以启发和感动。	The essay **excels** in focus and relevance. The writer depicts **Yu Wei**, a character chosen for her vivid spirit shown through qualities and actions. **Yu Wei** is presented as **academically and morally excellent, respectful, gentle, and intelligent**. Her friendship and assistance **touched** the writer, who reflects on **mutual support, valuing friendship, and hope for the future**. The piece stays on point with sincere emotions, inspiring and moving readers.
GPT-4 (b)	这篇作文的中心与切题表现**优秀**。作者选择了一位**品学兼优、乐于助人**的女孩作为写作对象，通过描写她的品质和行为，表达了对她的敬佩和感激之情。同时，文章也强调了**友谊**的重要性，以及**互帮互助**的精神。从切题的角度来看，这篇作文**完全符合题目要求**，以一个具体的人物为例，深入挖掘其精神品质和代表的风气，并**表达了作者的情感倾向**。因此，我认为这篇作文的中心与切题方面**都做得很好**。	The essay **excels** in focus and relevance. The author chose a girl who **excels academically and is eager to help others** as the subject for the essay, using descriptions of her qualities and actions to convey admiration and gratitude. The piece also highlights **the significance of friendship and the spirit of mutual assistance**. From the perspective of relevance, the essay **fully meets** the requirements, using a specific individual to delve into her moral fiber and the ethos she represents while **expressing the author's emotional stance**. Hence, I believe this essay **excels in both focus and relevance**.
Teacher (c)	这篇作文在中心和切题方面**不合格**。文章的题目是《这样的人让我感动》，应该侧重于写让作者感动的人的事迹，进而表现这个人的性格、品质等特征。本文却**从友谊说起，讲二人之间是如何成为朋友，发展友情的，与题目要求不符**。建议作者在写作前**仔细审题，围绕中心组织事实材料**。	The essay **fails** to meet the requirements in relevance and focus. The title is "Such People Touch Me" which should concentrate on the deeds of the person who moved the author, thereby showcasing their personality and qualities. **However, the essay starts with friendship, discussing how two people become friends and develop their relationship, deviating from the topic.** The author is advised **to carefully review the question before writing and organize facts around the central theme.**

Figure 2: An example of topic relevance comment for one grade 7 student essay titled *Such People Move Me* based on EduChat, GPT-4 and a Chinese primary school teacher. More details of this example are shown in Appendix B. Texts highlighted in **red** indicate misinformation or incorrect advice, whereas texts highlighted in **blue** provide accurate feedback for students. Specifically, in part (a), EduChat incorrectly assesses the level as "excel" and fabricates a character named Yu Wei that does not exist; in part (b), GPT-4 inaccurately evaluates the essay's level as "fully meets" and fails to identify the specific "author's emotional stance"; in part (c), teacher-written comment correctly points out the topic derivation of this essay, explains why, and provides detailed suggestions for refinement.

paucity of research within Chinese essay evaluation. Zhang et al. (2022) represents a limited foray into this area, concentrating on comment generation for Chinese narrative essays at the sentence and paragraph levels for student cohorts in the fifth grade and beyond. However, it neglects the diverse requirements posed by various writing tasks and is confined to a singular literary genre. Recently, Large Language Models (LLMs) have been exhibiting considerable promise across various applications (Zhao et al., 2023). Nonetheless, these models, including GPT-4 [2] and domain-specific iterations like EduChat (Dan et al., 2023), often make incorrect judgements regarding the topic relevance and generate overly generic or even misleading feedback when tasked with assessing topic relevance of essays, which is related to *hallucination* of LLMs, as shown in Figure 2.

To address these issues, we introduce a novel dataset **TOREE** (**To**pic **Re**levance **E**valuation) fo-

cusing on evaluating topic relevance of essays written by Chinese primary and middle school students. To the best of our knowledge, TOREE offers three primary advantages over existing works in this field. Firstly, TOREE is **the first attempt** to evaluate Chinese essays based on their adherence to topic relevance, which is more comprehensive, considering both the objectives of the writing assignment and the substance of the essay, rather than focusing solely on content. Secondly, TOREE encompasses a broad range of participants, specifically targeting students in the compulsory education stage of China, which includes **grades 3 through 9**. This span captures **13** distinct literary genres and covers **74** diverse essay prompts, thereby providing insights into student writing abilities across different styles and levels. Thirdly, TOREE is constructed as a **comparison dataset** that is apt for contrasting and evaluating models through both Supervised Fine-tuning (SFT) and Preference Learning (PL) (Rafailov et al., 2023) like Reinforcement Learning from Human Feedback (RLHF), making it highly

---

[2] https://chat.openai.com

5750

versatile for AES, AECG, and other various applications. Two tasks based on TOREE are designed for a comprehensive evaluation: **Topic Relevance Recognition** and **Topic Relevance Comment Generation**. Moreover, we propose a two-step method combining SFT and PL to elicit abilities of open-source LLMs to identify topic relevance level and generate specific and constructive comments. Experimental results suggest that our method effectively enhances models' abilities on two tasks, improving both automatic and human evaluations across 4 diverse LLMs. To sum up, our contributions can be summarized as follows.

- We introduce TOREE, a novel comprehensive dataset for evaluating topic relevance of essays, which is beneficial for AES, AECG, and other applications.

- TOREE can be utilized through both SFT and PL. Open-source LLMs that have been trained with our method demonstrate a substantially improved ability to both identify topic relevance level and generate specific feedback.

- TOREE is a valuable resource for further exploration from the perspective of PL.

2 Related Work

2.1 Automatic Essay Scoring

The goal of AES is to predict holistic or trait-specific rankings, or numerical scores to essays by analyzing a range of textual features, such as coherence, grammatical correctness, and expressive clarity (Jiang et al., 2023; Song et al., 2020; Farag et al., 2018). Earlier research has utilized traditional Machine Learning (ML) methods like Support Vector Machine (SVM) (Jin et al., 2018). Deep Learning based methods include using CNN (Dong et al., 2017), LSTM (Tay et al., 2018) to predict scores, and more recently, focus has shifted to small language models (SLMs) like BERT (Yang et al., 2020; Mayfield and Black, 2020; Cao et al., 2020; Uto et al., 2020) and more complicated attention-based design (He et al., 2022) for this purpose. Scoring certain essay traits seeks to attribute different scores to various characteristics of an essay such as clarity (Ke et al., 2019) and style (Mathias and Bhattacharyya, 2018). However, AES systems mainly provide convenience for educators but offer little to no feedback for students to improve their essays. In this study, our objective is not only predicting scores, but also generating constructive topic relevance feedback with comprehensive advice on how to enhance essays to revolve around a certain them and more closely match the requirements of given essay prompts.

2.2 Automatic Essay Comment Generation

Despite the myriad of existing studies (Raheja et al., 2023; Dwivedi-Yu et al., 2022; Shi et al., 2022) and the availability of tools designed to aid writers, such as Grammarly [3] and current LLMs, these resources predominantly offer superficial assistance, which are adept at addressing grammatical inaccuracies or enhancing stylistic elements but fall short in enhancing the topic relevance of an essay. AECG targets at generating effective feedback for different students to purposefully revise their essays, thereby improving writing skills. To date, there has been a scarcity of research in this area, owing to its inherent complexity and the challenges involved. Zhang et al. (2022) leveraged a plan-based approach to generate feedback using the T5 model (Raffel et al., 2020), which was designed to evaluate Chinese narrative essays by pinpointing their strengths and weaknesses. Recent advancements in LLMs have led to the emergence of education-focused LLMs which hold significant potential for educational applications, such as EduChat (Dan et al., 2023). However, the feedback provided by such models is often too general to aid students in specifically improving their essays or may even include flase information, potentially misleading students. Currently, ASAP [4] and TOEFL11 (Blanchard et al., 2013) are two most frequently used public datasets within the Automatic Essay Evaluation field. However, both datasets merely contain 8 essay prompts. In contrast, our research examines essay evaluations for a broader range of students, specifically those in the entire compulsory education stage, which includes **grades 3 through 9**, spans **13** different literary genres and covers **74** diverse essay prompts. Furthermore, we aim to provide more detailed feedback that can offer students clear, actionable advice on how to enhance their essays in a manner tailored to each individual.

---

[3] https://www.grammarly.com
[4] https://www.kaggle.com/c/asap-aes/data

3 The TOREE Dataset

3.1 Task Definition

We designed two tasks according to the levels of comprehension and generation. The objective of Task 1 is to identify the topic relevance level of an essay, while the goal of Task 2 is to generate specific and detailed comments for students.

Task 1: Topic Relevance Recognition   Task 1 is designed as a 5-category classification task, mapping the topic relevance of an essay to one of five levels: *Excellent*, *Good*, *Fair*, *Passable*, and *Failed*. The specific criteria for these levels were designed and discussed by professional annotators with extensive Chinese teaching experience in our annotation team, with details provided in Appendix C.1.

Task 2: Topic Relevance Comment Generation
The focus of Task 2 is pointing out the areas of the essay that effectively stay on topic, while also offering specific and detailed suggestions for revising essays to enhance relevance to the essay prompt, such as assistance with modifying the title, adjusting the narrative order, rewriting certain parts, etc.

Data Format   Taking into account the use of SFT to foster the ability to generate comments, as well as the utilization of PL to better align with the preferences of educators, the structure of TOREE is articulated as a five-fold combination, comprising (*instruction, essay, topic_relevance_level, chosen_comment, rejected_comment*). Within the *instruction* field, the essay prompt and the grade of essay's author are provided. The *essay* field includes both the title of the essay and its body content and the *topic_relevance_level* field is the annotated topic relevance level defined in Task 1. Furthermore, the *chosen_comment* and *rejected_comment* form a pair of preference data. More details about the preference are provided in Appendix C.2.

3.2 Dataset Construction

Essay Collection   We gathered 730 original essays from the regular writing exercises of primary and middle school pupils and sorted by genre and grade. Through the implementation of rule-based filtering operations, we ensured the quality of essays to some extent. As a result, we ended up with a collection of 683 essays, with the distributions depicted in Figure 3. More details of our filtering rules are shown in Appendix C.3.

	Train	Valid	Test
# Essays	544	67	72
# Grades	7	7	7
# Genres	13	13	13
Avg. Title Len	5.79	5.69	6.03
Avg. Content Len	488.80	491.39	538.07
Avg. Chosen Comment Len	150.33	148.96	146.38
Avg. Rejected Comment Len	158.57	156.75	160.36
Topic Relevance Level Ratio			
Excellent	41.91%	40.30%	50.00%
Good	43.01%	40.30%	31.94%
Fair	6.99%	8.96%	9.72%
Passable	2.02%	1.76%	1.39%
Failed	6.07%	8.68%	6.55%

Table 1: Overall statistics of TOREE. *Len* denotes *Length*, which is computed by counting the number of Chinese characters. The Topic Relevance Level Ratio denotes the ratio of essays labeled in the certain level.

LLM Generation   We selected Spark[5] to generate topic relevance comments because of its relatively notable strengths in the field of Chinese education. In particular, we provided it with an essay prompt along with specific requirements for assessing topic relevance, which is shown in Appendix D. We chose Spark-v3 for generation and further details are provided in Section 5.3.2. Although Spark is adept at evaluating essays compared to other LLMs, the comments it produces still face hallucination problems and cannot be directly used for guiding students. Therefore, we put these comments in the *rejected_comment* field.

Human Annotation   To start, we selected our potential annotators from a group of 20 individuals majoring in Chinese studies, ensuring they had at least an undergraduate degree and skills in reading and writing Chinese text. Moreover, following HaluEval (Li et al., 2023), these candidates were asked to annotate a few test samples for pre-annotation. We assessed their reliability using Fleiss's Kappa κ for Task 1 and BERTscore (Zhang et al., 2019) for Task 2. We chose the top 6 performers as our final annotators. For Task 1, all annotators labeled the topic relevance level and we obtained $\kappa = 0.812$ ($0.80 \leq \kappa \leq 1.00$), showing a perfect agreement. For Task 2, within the annotation team, we divided the responsibilities, assigning three annotators to evaluate essays from primary school students in grades 3 to 6 and another three to assess essays from middle school students in grades 7 to 9. Consequently, we have
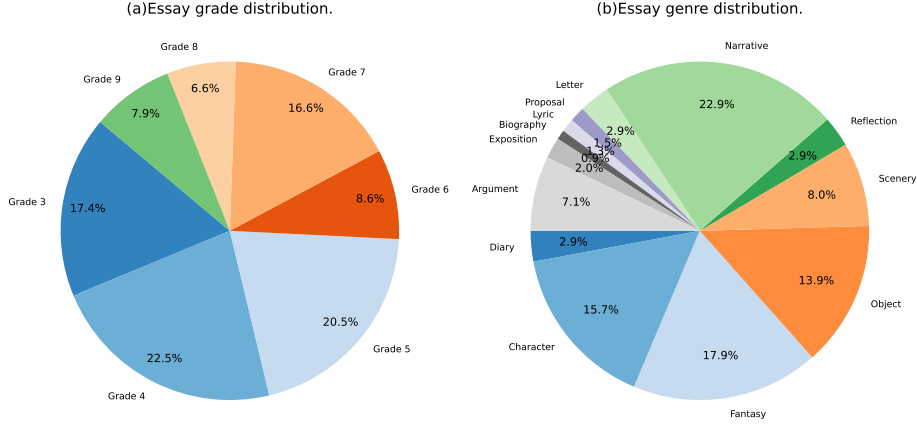
---

[5]https://xinghuo.xfyun.cn/

Figure 3: The distribution of TOREE. TOREE comprises 683 essays, each accompanied by a pair of chosen and rejected topic relevance comments. Part (a) displays the distribution of essays according to educational level, encompassing seven grades of compulsory education stage in China. The omission of the first and second grades is attributed to the educational focus at these stages on acquiring Chinese characters and sentence construction, rather than essay composition. Part (b) presents the distribution of essays by literary genre, including 13 unique genres.

included the superior comments made by our annotators in the *chosen comment* field. More details of annotation are provided in Appendix C.4.

3.3 Dataset Statistics

As depicted in Figure 3, TOREE encompasses a comprehensive range of essay topics typically assigned to students within the seven grades of compulsory education stage in China, which includes 683 essays, spanning across **13** different literal genres and **74** diverse prompts. Notably, the curriculum for the first and second grades of primary school mainly focuses on learning characters and writing sentences, rather than writing essays. Detailed descriptions of genres and prompts are shown in Appendix C.5. Additionally, we divide TOREE into a training, validation, and testing set in a proportion of 8:1:1, with the details in Table 1.

3.4 Ethical Issues

We paid all the data annotators for their contributions. Additionally, we obtained clear consent from the essay authors and their legal guardians to use the essays for the purposes of labeling and sharing the work publicly. To protect the students' privacy, we have removed any personal information such as names and addresses from the essays in our dataset.

4 Method

We propose a two-step method for recognizing topic relevance levels and generating corresponding topic relevance comments in a joint format.

4.1 Step 1: Supervised Fine-tuning

In the first step, we apply Supervised Fine-tuning for eliciting LLMs' abilities to recognize levels and comment on the topic relevance aspect of an essay. Specifically, the training objective is to minimize the standard cross-entropy loss for auto-regressive generation, defined as:

$$\mathcal{L}_{\text{SFT}} = -\frac{1}{B}\sum_{i=1}^{B}\log p_i(y_i|x_i) \qquad (1)$$

where B denotes the batch size, y_i denotes the concatenation $level \oplus chosen\_comment$, and x_i denotes the concatenation $instruction \oplus essay$.

4.2 Step 2: Preference Learning

We incorporate DPO (Rafailov et al., 2023) for further alignment with human comments. Specifically, we utilize LLMs' intrinsic abilities to learn from preference comments, similar to RLHF. The training objective in this step is defined as:

$$\mathcal{L}_{\text{DPO}} = -\mathbb{E}\left[\log\sigma\left(\beta\log\frac{\pi_\theta(y_i^c|x_i)}{\pi_{\text{ref}}(y_i^c|x_i)} -\beta\log\frac{\pi_\theta(y_i^r|x_i)}{\pi_{\text{ref}}(y_i^r|x_i)}\right)\right] \quad (2)$$

where $\mathbb{E}$ denotes the expectation over a batch, σ denotes Sigmoid function, π_θ is the model undergoing training, π_{ref} is the frozen reference model (trained after SFT) and β is a hyperparameter controlling the deviation from the reference model.

Similar to Equation 1, x_i is the concatenation $instruction \oplus essay$, y_i^c is the chosen comment concatenation $level \oplus chosen\_comment$, and y_i^r is the rejected comment concatenation $level \oplus rejected\_comment$ in the data point.

5 Experiment

5.1 Experimental Setup

Models We consider a broad range of methods in Automatic Essay Evaluation, including ML methods, SLMs, and LLMs. For ML methods, we selected RankSVM (Jin et al., 2018) and Random-Forest for Task 1. For SLMs, we selected BERT (Mayfield and Black, 2020), XLNet (Yang et al., 2019), and RoBERTa (Liu et al., 2019) for Task 1. Regarding LLMs, we selected both open-source and closed-source LLMs that are currently supportive of Chinese language and are prevalently utilized in current research for both Task 1 and Task 2. For open-source models, we utilized Baichuan2-7B (Yang et al., 2023), ChatGLM3-6B (Du et al., 2022), Chinese-Alpaca2-7B (Cui et al., 2023), Qwen-7B (Bai et al., 2023), and EduChat (Dan et al., 2023). In terms of closed-source models, our selection comprised of ChatGPT (*gpt-3.5-turbo-0613*), GPT-4 (*gpt-4-turbo-preview*), Claude (*claude-3-sonnet-20240229*) and Spark.

Evaluation Metrics For Task 1, the metrics utilized to evaluate performance are **Precision**, **Recall**, and **Weighted-F1**, while for Task 2, we use both automatic evaluation scores and human evaluation because of the complexities associated with this task. In terms of automatic evaluation scores, we select **BLEU**, **ROUGE**, and **BERTscore** (Zhang et al., 2019) for gauging the similarity in terms of both syntax and meaning between the generated comment and the reference comment (the *chosen_comment* field in a data point). For human evaluation, we incorporate the expertise of three annotators in our team to compare generated comments from the perspective of real-world educational applications. They evaluate pairs of comments for identical compositions, rendering a result of *Win/Tie/Lose* for each pair and the final result is obtained through majority vote.

5.2 Implementation Details

For testing, we compute Precision, Recall, and Weighted-F1 for ML and SLM methods under zero-shot setting on Task 1 and compute automatic evaluation scores for LLMs under few-shot setting on Task 1 and Task 2. Notably, given that context length of LLMs is limited and each essay is relatively lengthy, we chose 2 shots for test. Regarding SFT and PL training, we adopted LoRA (Hu et al., 2021) with the LoRA rank of 8 and the dropout rate of 0.1 across all 3 epochs. All SFT runs were performed on a single NVIDIA RTX 3090 GPU, while DPO runs were executed on a single A100 GPU. Given the nature of the TOREE (shown in Table 1), we limited the maximum length to 1024 tokens. Training configurations include the learning rate of 5×10^{-5}, the batch size of 2, and the employment of gradient accumulation with a step size of 2 across three epochs. In addition, we implemented a Cosine learning rate scheduler without the inclusion of warm-up steps and enable mixed precision training (fp16) to enhance efficiency.

5.3 Results and Analysis

5.3.1 Topic Relevance Recognition

The complete results are shown in Appendix E.

Poor Performance of Current LLMs As shown in left part of Table 2, the results reveal that both closed-source and open-source LLMs exhibit deficiencies in accurately assessing the topical relevance of essays, a performance lag evident when compared to SLMs and traditional ML methods. For example, the state-of-the-art GPT-4 achieves 36.29 and 43.07 Weighted-F1 under 2-shot and zero-shot settings, which is 4.36 lower than BERT. Furthermore, performance metrics for other open-source LLMs consistently fall below a Weighted-F1 score of 30. In contrast, our method markedly enhances the ability of LLMs to determine the topic relevance of essays. Specifically, the Weighted-F1 score for the Baichuan2-7B significantly improved from 20.31 to 42.86. This score not only demonstrates a substantial increase but also exceeds the GPT-4 under zero-shot scenario by 6.57 points.

Effect of Few-shot Prompting The effect of Few-shot Prompting on various models differs on Task 1. For closed-source LLMs except Spark-v2, few-shot prompting brings Weighted-F1 score **gains** up to 15.17 (Claude), while for open-source LLMs except EduChat, few-shot prompting brings huge performance **drop** on Weighted-F1 score up to 21.19 (Qwen-7B). As a common belief, few-shot prompting can enhance LLMs' abilities on various downstream tasks by providing additional information that helps LLMs make decisions. We

Type	Model	Task 1			Task 2				
		Precision	Recall	Weighted-F1	B-4	R-1	R-2	R-L	BERTscore
ML	RankSVM	51.39	46.27	<u>47.01</u>	–	–	–	–	–
	RandomForest	41.67	43.75	41.67	–	–	–	–	–
SLM	BERT	46.35	**50.00**	**47.43**	–	–	–	–	–
	XLNet	38.51	43.06	39.65	–	–	–	–	–
	RoBERTa	44.52	43.06	43.18	–	–	–	–	–
Closed-source LLM	Claude$_{0-shot}$	27.15	29.17	27.33	28.21	41.19	13.34	21.51	68.79
	Claude$_{2-shot}$	48.43	43.06	42.50	28.09	41.40	13.43	21.74	68.82
	ChatGPT$_{0-shot}$	48.16	25.00	32.86	25.78	37.44	13.65	22.68	67.39
	ChatGPT$_{2-shot}$	46.23	31.94	37.63	26.05	36.82	13.50	22.67	67.18
	GPT-4$_{0-shot}$	33.06	**50.00**	36.29	32.28	43.43	15.64	24.44	69.86
	GPT-4$_{2-shot}$	<u>51.43</u>	37.50	43.07	26.17	38.18	13.49	22.58	67.36
	Spark-v1$_{0-shot}$	21.42	20.83	21.12	32.44	<u>47.20</u>	19.01	25.99	69.64
	Spark-v1$_{2-shot}$	29.01	26.39	25.62	27.65	42.37	18.09	24.61	67.62
	Spark-v2$_{0-shot}$	25.00	<u>48.61</u>	33.02	<u>35.32</u>	47.54	19.54	27.92	70.75
	Spark-v2$_{2-shot}$	24.26	45.83	31.73	34.02	45.59	20.03	27.26	70.22
	Spark-v3$_{0-shot}$	37.42	<u>48.61</u>	37.64	**35.70**	47.04	19.73	28.46	71.10
	Spark-v3$_{2-shot}$	39.52	47.22	38.68	32.89	46.32	20.38	27.89	70.28
Open-source LLM	EduChat$_{0-shot}$	42.00	18.06	23.06	22.71	39.36	13.92	22.56	68.08
	EduChat$_{2-shot}$	31.93	22.22	25.95	24.43	36.23	13.05	22.04	66.73
	Chinese-Alpaca2-7B$_{0-shot}$	35.71	6.94	11.63	4.78	20.75	4.13	16.68	62.73
	Chinese-Alpaca2-7B$_{2-shot}$	19.23	6.94	10.20	1.06	13.58	0.87	12.77	57.53
	ChatGLM3-6B$_{0-shot}$	29.41	13.89	18.87	4.78	17.49	4.35	12.31	59.27
	ChatGLM3-6B$_{2-shot}$	21.43	8.33	12.00	1.10	12.16	0.82	12.50	56.76
	Qwen-7B$_{0-shot}$	37.73	27.78	30.82	10.45	31.46	8.59	23.27	69.27
	Qwen-7B$_{2-shot}$	36.94	5.56	9.63	1.02	12.27	0.66	13.06	56.16
	Baichuan2-7B$_{2-shot}$	50.00	2.78	5.26	1.40	13.99	1.04	13.18	57.46
	Baichuan2-7B$_{0-shot}$	**61.94**	13.89	20.31	4.79	17.81	3.62	13.89	60.24
	w/ SFT	40.17	**50.00**	42.23	25.96	44.33	<u>24.13</u>	<u>38.58</u>	<u>75.86</u>
	w/ SFT + DPO	44.74	43.06	42.86	25.98	46.56	**25.28**	**40.15**	**76.26**

Table 2: Automatic evaluation results on TOREE test set. For the sake of brevity, for LLMs only the result of Baichuan2-7B trained using our method is demonstrated here, with the comprehensive results available in Appendix E. Best results are highlighted in **bold**, and second best results are <u>underlined</u>.

hypothesize that closed-source LLMs and EduChat (which is specifically pre-trained on Chinese textbook corpus from education domain) have stronger Chinese comprehension capabilities, thereby enhancing Weighted-F1 score on Task 1.

Error Analysis To delve deeper into the predictions of LLMs and better understand our method, we present confusion matrices on Task 1 in Figure 4. Notably, GPT-4 exhibits a strong tendency to predict the test data into **higher** levels, with almost all its predictions concentrated in the top three categories. We hypothesize that this problem may be a form of *sycophancy* (Sharma et al., 2023) in the domain of essay assessment, a specific type of *hallucination* (Zhang et al., 2023), which can lead to an overly optimistic evaluation of student essays. Moreover, the performance of Baichuan2-7B is remarkably poor, failing to successfully predict any of the five categories in the vast majority of cases. This suggests that current vanilla open-

source LLMs may have significant limitations or deficiencies in handling task 1. Compared to GPT-4, Baichuan2-7B-SFT shows a slight mitigation of sycophancy, but the improvement is not substantial. On the other hand, Baichuan2-7B-SFT-DPO effectively alleviates the sycophancy issue, resulting in a higher number of correct predictions and a more balanced distribution of predicted labels across categories, with fewer overly optimistic predictions towards higher grades. More explanations of error analysis on Task 1 are provided in Appendix F.

5.3.2 Topic Relevance Comment Generation

The right part of Table 2 presents automatic evaluation scores of evaluated LLMs on Task 2.

Automatic Evaluation Scores Drop of Few-shot Prompting Unlike Task 1, Few-shot prompting leads to a decline in all automatic evaluation scores on both open-source and closed-source LLMs across almost all models. Similarly, we surmise that models' ability to process long texts and

GPT-4

True \ Pred	Excellent	Good	Fair	Passable	Failed	None
Excellent	29	3	0	0	0	0
Good	22	1	0	0	0	0
Fair	6	0	1	0	0	0
Passable	4	1	0	0	0	0
Failed	3	1	0	0	1	0
None	0	0	0	0	0	0

Baichuan2-7B

True \ Pred	Excellent	Good	Fair	Passable	Failed	None
Excellent	5	0	0	0	0	27
Good	0	1	0	0	0	22
Fair	0	0	0	0	0	7
Passable	0	0	1	4	0	0
Failed	0	0	0	0	0	5
None	0	0	0	0	0	0

Baichuan2-7B-SFT

True \ Pred	Excellent	Good	Fair	Passable	Failed	None
Excellent	12	18	1	1	0	0
Good	8	13	2	0	0	0
Fair	5	1	1	0	0	0
Passable	1	0	0	4	0	0
Failed	4	0	0	0	1	0
None	0	0	0	0	0	0

Baichuan2-7B-SFT-DPO

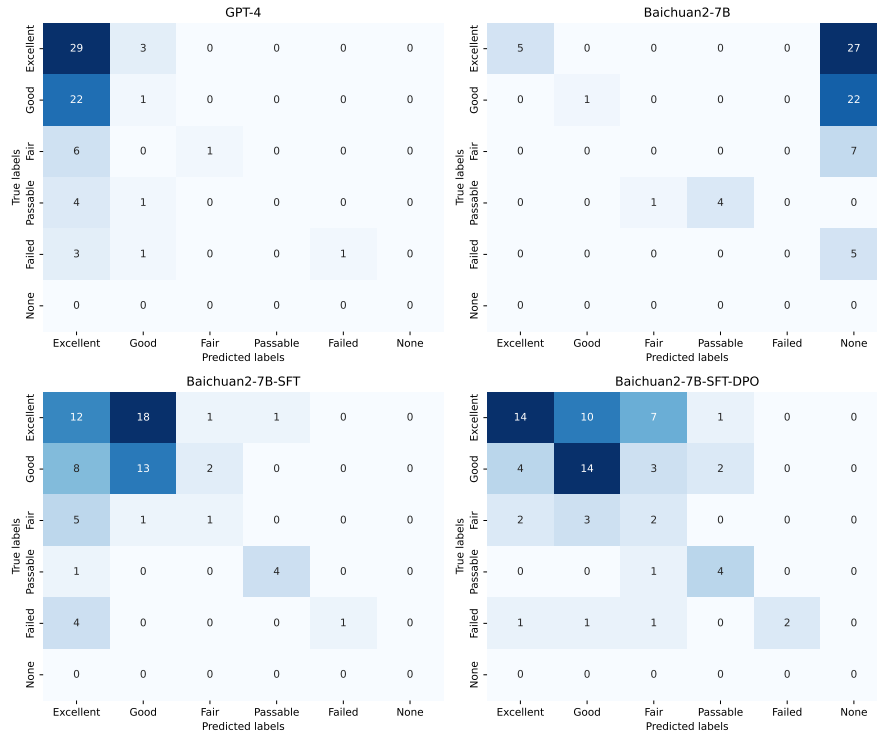| True \ Pred	Excellent	Good	Fair	Passable	Failed	None
Excellent	14	10	7	1	0	0
Good	4	14	3	2	0	0
Fair	2	3	2	0	0	0
Passable	0	0	1	4	0	0
Failed	1	1	1	0	2	0
None	0	0	0	0	0	0

Figure 4: Confusion matrices of GPT-4, Baichuan2-7B, Baichuan2-7B-SFT, and Baichuan2-7B-SFT-DPO. *None* denotes model fails to predict a label in five categories.

LLM	Preference Data	BLEU-4	ROUGE-L	BERTscore
Baichuan2-7B	h-v1	**25.98**	**40.15**	**76.26**
	h-v2	25.33	39.61	76.04
	h-v3	25.25	38.99	75.94
	h-v1,h-v2,h-v3	22.71	36.25	74.56
Chinese-Alpaca2-7B	h-v1	**25.33**	**39.69**	**76.04**
	h-v2	23.58	37.78	74.80
	h-v3	22.51	34.64	74.09
	h-v1,h-v2,h-v3	20.07	32.68	71.56
ChatGLM3-6B	h-v1	**24.18**	**37.26**	**75.07**
	h-v2	24.10	36.91	75.05
	h-v3	23.66	36.27	75.02
	h-v1,h-v2,h-v3	22.84	36.04	74.46
Qwen-7B	h-v1	**28.17**	**42.31**	**76.85**
	h-v2	27.73	41.85	76.73
	h-v3	26.84	40.54	76.54
	h-v1,h-v2,h-v3	25.47	38.59	73.19

Table 3: Results on Task 2 of models trained with different preference data in Step 2. *h-v1* denotes preference data composed of human and Spark-v1, with other representations following in a similar manner.

LLM	Comparison	Human Evaluation		
		Win	Tie	Lose
Baichuan2-7B	SFT vs GPT-4	72.2	25.0	2.8
	DPO vs GPT-4	80.6	16.7	2.7
	DPO vs SFT	54.0	14.0	32.0
Chinese-Alpaca2-7B	SFT vs GPT-4	66.7	25.0	8.3
	DPO vs GPT-4	75.0	16.7	8.3
	DPO vs SFT	58.3	13.9	27.8
ChatGLM3-6B	SFT vs GPT-4	70.8	23.6	5.6
	DPO vs GPT-4	79.2	16.7	4.1
	DPO vs SFT	56.9	12.5	31.6
Qwen-7B	SFT vs GPT-4	69.4	23.6	7.0
	DPO vs GPT-4	76.4	16.7	6.9
	DPO vs SFT	55.6	13.9	31.5

Table 4: Human evaluation results on TOREE test set.

comprehend essays is limited and Task 2 is too challenging. Consequently, these models struggle to differentiate between the essay to be evaluated and the demonstration essays provided, leading to some of the generated comments being influenced by the example essays and commenting on the wrong target. We believe that the example essays provided as in-context demonstrations introduced a certain degree of "noise" in this context.

In-Depth Analysis of DPO As shown in Equation 2 in Step 2, LLMs further learn to align with human through preference learning. We delved into the impact of varying preference data on the performance on Task 2. Specifically, in light of the results presented in Table 2, the capability of three versions of Spark to generate topic relevance comments has progressively improved. Furthermore, considering the cost associated with closed-source model APIs, we constructed three distinct preference datasets using these three models. It can

LLMs	Comments
GPT-4	In terms of addressing the theme, your composition excels with a commendable adherence to the theme "Me in Ten Years". The use of contrast, by comparing with "myself" ten years prior, effectively accentuates the theme. However, while the composition vividly imagines future life, it lacks detail in terms of the specific professional undertakings and psychological disposition in the future. Expansion in these areas could enrich the essay further. Overall, your composition is thematic, adherent to the prompt, and stands out as an exemplary piece with depth and innovation.
Baichuan2-7B-SFT	The essay on "Me in Ten Years" meets the criteria for relevance and central theme adequately. By employing a rich imagination, the piece portrays the detailed aspects of life ten years hence, not only highlighting the characteristics of technological advancements but also illustrating the author's adaptability. The imagination is realistic, yet it falls short of thoroughly exploring personal content related to the author.
Baichuan2-7B-SFT-DPO	The essay titled "Me in Ten Years" does not fulfill the criteria for relevance and central theme. However, the piece lacks a depiction of the social context, failing to provide a detailed representation of societal evolution or personal psychological shifts, resulting in a narrative that feels somewhat superficial. It is suggested that the author review the essay prompt again thoroughly and ensure your essay focuses on the exploration of characters.
Teacher	The essay "Me in Ten Years" fails to meet the expectations in terms of relevance and central theme. Although the author employs a vivid imagination to narrate the sequence of waking up in one's own house ten years in the future, using concrete elements like "sunglasses" and "dressing room" to depict changes in societal living, there is a significant omission in the description of personal characteristics and mindset. This lack of introspection regarding "myself" leaves the narration somewhat wanting. It is recommended that the author thoroughly review the essay prompt before proceeding to ensure that the narrative centers around the exploration of "myself".

Table 5: Case study on the essay *Me in Ten Years*. Texts highlighted in **red** indicate misinformation or incorrect advice, whereas texts highlighted in **blue** provide accurate feedback.

be inferred that the discrepancy between Spark-v3 comments and human annotations is the minimal, followed by Spark-v2, with Spark-v1 exhibiting the largest gap. The results are shown in Table 3. Our findings indicate that all models performed optimally in the scenario of *h-v1*, as evidenced by the highest scores of BLEU-4, ROUGE-L, and BERTscore. Furthermore, aggregating all preference data results in a decline in performance across all models, which may originate from the presence of multiple preferences for the same essay. Consequently, it can be inferred that the larger the disparity between the chosen comments and rejected comments within preference data, the more the model is able to learn, ultimately resulting in more effective comment. We believe this insight is helpful for both automatic essay evaluation and other domains involving preference learning.

Human Evaluation We conducted human evaluation experiments on comments generated by models after training with SFT and models after training with SFT and DPO, setting GPT-4 under 2-shot setting as baseline. As shown in Table 4, our method significantly enhances the model's capability to generate topic relevance comments, with outcomes that are markedly superior to GPT-4 from the perspective of real-world applications. For instance, Baichuan2-7B-SFT-DPO achieved a win rate of 80.6% against GPT-4, and the win rates for other models did not fall below approximately 70%. Furthermore, the ablation results for all models (DPO vs SFT) also demonstrate the efficacy of our two-

step approach: the generated comments are more aligned with human-written feedback and possess greater practical value.

6 Case Study

As shown in Table 5, regarding this **off-topic** essay, GPT-4 mistakenly considers its central theme and relevance to be *excellent*. Although it provides a detailed suggestion describing the author's specific work, the overall direction of the feedback is misleading to the student. In comparison, Baichuan2-7B-SFT offers recommendations instead of vague descriptions, while Baichuan2-7B-SFT-DPO identifies specific issues and provides executable suggestions (*"review the essay prompt again thoroughly and ensure your essay focuses on the exploration of characters"*), which is more closely resembling teacher comments.

7 Conclusion

In conclusion, this paper introduces TOREE, a novel comprehensive dataset for assessing the topic relevance of student essays, addressing a critical need in educational assessment. TOREE serves as a valuable resource for future research and development in the field of automatic essay evaluation, which can be further utilized through Supervised Fine-tuning or Preference Learning. Moreover, our two-step approach has demonstrated success in enhancing the performance of open-source Large Language Models in recognizing topic relevance and providing constructive feedback, presenting great value in real-world applications.

Limitations

The main limitation of the study is the relatively limited number of essays in the TOREE dataset, which is a consequence of the challenges associated with data acquisition. Obtaining a diverse and representative sample of student essays, particularly from the compulsory education stage, requires extensive collaboration with educational institutions and adherence to privacy regulations. Despite these difficulties, the dataset's quality and the tasks designed around it provide a solid foundation for future research and model development in the field of automatic essay evaluation.

Acknowledgment

References

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Daniel Blanchard, Joel Tetreault, Derrick Higgins, Aoife Cahill, and Martin Chodorow. 2013. Toefl11: A corpus of non-native english. *ETS Research Report Series*, 2013(2):i–15.

Yue Cao, Hanqi Jin, Xiaojun Wan, and Zhiwei Yu. 2020. Domain-adaptive neural automated essay scoring. In *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval*, pages 1011–1020.

Yiming Cui, Ziqing Yang, and Xin Yao. 2023. Efficient and effective text encoding for chinese llama and alpaca. *arXiv preprint arXiv:2304.08177*.

Yuhao Dan, Zhikai Lei, Yiyang Gu, Yong Li, Jianghao Yin, Jiaju Lin, Linhao Ye, Zhiyan Tie, Yougen Zhou, Yilei Wang, et al. 2023. Educhat: A large-scale language model-based chatbot system for intelligent education. *arXiv preprint arXiv:2308.02773*.

Fei Dong, Yue Zhang, and Jie Yang. 2017. Attention-based recurrent convolutional neural network for automatic essay scoring. In *Proceedings of the 21st conference on computational natural language learning (CoNLL 2017)*, pages 153–162.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335.

Jane Dwivedi-Yu, Timo Schick, Zhengbao Jiang, Maria Lomeli, Patrick Lewis, Gautier Izacard, Edouard Grave, Sebastian Riedel, and Fabio Petroni. 2022. Editeval: An instruction-based benchmark for text improvements. *arXiv preprint arXiv:2209.13331*.

Youmna Farag, Helen Yannakoudakis, and Ted Briscoe. 2018. Neural automated essay scoring and coherence modeling for adversarially crafted input. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 263–271, New Orleans, Louisiana. Association for Computational Linguistics.

Yaqiong He, Feng Jiang, Xiaomin Chu, and Peifeng Li. 2022. Automated chinese essay scoring from multiple traits. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3007–3016.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2021. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Zhiwei Jiang, Tianyi Gao, Yafeng Yin, Meng Liu, Hua Yu, Zifeng Cheng, and Qing Gu. 2023. Improving domain generalization for prompt-aware essay scoring via disentangled representation learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12456–12470, Toronto, Canada. Association for Computational Linguistics.

Cancan Jin, Ben He, Kai Hui, and Le Sun. 2018. Tdnn: a two-stage deep neural network for prompt-independent automated essay scoring. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1088–1097.

Zixuan Ke, Hrishikesh Inamdar, Hui Lin, and Vincent Ng. 2019. Give me more feedback ii: Annotating thesis strength and related attributes in student essays. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 3994–4004.

Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. Halueval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6449–6464.

Ming Liu, Yi Li, Weiwei Xu, and Li Liu. 2017. Automated essay feedback generation and its impact on revision. *IEEE Transactions on Learning Technologies*, 10(4):502–513.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Yuanchao Liu, Jiawei Han, Alexander Sboev, and Ilya Makarov. 2023. Geef: A neural network model for automatic essay feedback generation by integrating writing skills assessment. *Expert Systems with Applications*, page 123043.

Sandeep Mathias and Pushpak Bhattacharyya. 2018. Thank "goodness"! a way to measure style in student essays. In *Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications*, pages 35–41.

Elijah Mayfield and Alan W Black. 2020. Should you fine-tune bert for automated essay scoring? In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 151–162.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1).

Vipul Raheja, Dhruv Kumar, Ryan Koo, and Dongyeop Kang. 2023. Coedit: Text editing by task-specific instruction tuning. *arXiv preprint arXiv:2305.09857*.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R Johnston, et al. 2023. Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548*.

Ning Shi, Bin Tang, Bo Yuan, Longtao Huang, Yewen Pu, Jie Fu, and Zhouhan Lin. 2022. Text editing as imitation game. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1583–1594, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Wei Song, Ziyao Song, Lizhen Liu, and Ruiji Fu. 2020. Hierarchical multi-task learning for organization evaluation of argumentative student essays. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 3875–3881. International Joint Conferences on Artificial Intelligence Organization. Main track.

Yi Tay, Minh Phan, Luu Anh Tuan, and Siu Cheung Hui. 2018. Skipflow: Incorporating neural coherence features for end-to-end automatic text scoring. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Masaki Uto, Yikuan Xie, and Maomi Ueno. 2020. Neural automated essay scoring incorporating hand-crafted features. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6077–6088.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, et al. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

Ruosong Yang, Jiannong Cao, Zhiyuan Wen, Youzheng Wu, and Xiaodong He. 2020. Enhancing automated essay scoring performance via fine-tuning pre-trained language models with combination of regression and ranking. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1560–1569.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. 2023. Siren's song in the ai ocean: A survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*.

Zhexin Zhang, Jian Guan, Guowei Xu, Yixiang Tian, and Minlie Huang. 2022. Automatic comment generation for Chinese student narrative essays. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 214–223, Abu Dhabi, UAE. Association for Computational Linguistics.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

A Topic and Topic Relevance

For an essay, the *topic* refers to whether the essay has a clear central focus, without considering the relevance of that focus to the prompt requirements. In contrast, the *topic relevance* of an essay refers to whether the essay's central focus aligns with the prompt requirements, building upon having a central focus in the first place. In other words, previous work only focused on an essay's central topic itself, whereas TOREE focuses on the alignment between an essay's central topic and the prompt requirements (especially when the prompt constrains the genre of the essay). For example, an essay prompted to describe a scene would be considered **off-topic/off-prompt** even if it had excellent character descriptions. Given this, our work is the first to propose focusing on topic relevance when evaluating Chinese essays. Our primary motivation is to help students **avoid writing essays that are irrelevant to the prompt**, bridging the gap by evaluating topic relevance, aiming to improve the overall quality and responsiveness of student essays.

B Details of Example

Full details of the example shown in Section 1 are shown in Figure 5.

C More Details of TOREE

C.1 Topic Relevance Level

Details of topic relevance level in Task 1 are shown in Table 6. This division is discussed and determined by professional annotators for the following two primary reasons.

Correspondence with the Five-Level Evaluation Standard in Compulsory Education Stage This grading system aligns with the existing five-tier assessment criteria in the Chinese Language curriculum during the compulsory education phase. Teachers and students are more familiar with this structure, thereby enhancing their understanding and acceptance of the evaluation process.

Fine-Grained Assessment for Enhanced Feedback and Improvement The classification into five distinct levels offers a more nuanced approach to grading, allowing teachers to provide more detailed and accurate assessments of student work. In turn, students can utilize these specific grades to pinpoint areas for improvement and refine their

essays accordingly, leading to a clearer focus on topic relevance and overall writing quality.

C.2 Preference in Comments

Specifically, to meet the practical needs of Chinese language teachers in primary and middle schools and to help students improve their composition skills, we selected the preference in the TOREE that can be summarized as: whether the comment is specific and on-topic/on-prompt, whether it is targeted, and whether it provides feasible and executable suggestions for revision. For a pair of comments, we denote the one that better aligns with the above preference as the *chosen_comment*, and the one that deviates from the preference as the *rejected_comment*. Almost all comments generated by current state-of-the-art large language models, including those in the education domain such as Spark and EduChat, suffer from being too generic and vague, lacking specificity, and even exhibiting hallucinations, rendering them inapplicable for directly guiding students' composition revisions. Therefore, we treat comments annotated by human experts as *chosen_comment* and model-generated comments as *rejected_comment*. A specific example is shown in Table 8.

C.3 Essay Filtering

We adopted a two-fold filtering strategy. Firstly, we removed all essays manifesting substantial discrepancies due to the OCR processing errors, which significantly differ from the original ones. Subsequently, we removed all compositions exceeding 700 words or consisting of fewer than 200 words, which fail to satisfy the fundamental prerequisites of essay assignments.

C.4 Annotation

Annotating the topic relevance level and comment is a very challenging task, which requires good Chinese reading comprehension skills and using encouraging words to identify both the strong and weak points of an essay, as well as to provide constructive feedback for enhancing topic relevance. A screenshot of the annotation system is shown in Figure 6.

Team Composition and Selection The annotation team consisted of 6 selected individuals with backgrounds in Chinese language studies. These annotators were chosen based on their performance in a pre-annotation test, ensuring they possessed

题目：生活中我们会遇到各种各样的人，有的让你尊敬，有的让你佩服，有的让你感动，有的让你叹息。以《这样的人让我____》为题，写一篇作文。不少于500字。"这样的人"可以是你熟悉的人，比如你的同学、邻居，也可以是陌生的人，可以是某个具体的人，也可以是某一类人。"这样的人"应该是**具有某种精神品质**，或**代表某种风气**的人。要**着眼于个性、品质**去描写，写出你的**情感倾向**。

这样的人让我**感动**

如果说友谊是一颗常青树那么，浇灌它的必定是出自心田的清泉；如果说友谊是一朵开不败的鲜花，那么，昭耀它的必定是心中升起的太阳。初一新开学，几次摸底考，就有一个女孩尤唯特殊，常常考试第一名，尊敬师长，一个文静且聪慧的女孩，就是这样一个品学兼优的人，却被我开朗的性格所打动，与我成为了好朋友。我的成绩不太理想，算的上是班级的中后好，时常就感到自卑，我向她吐露心声。也许在那一刻，她决定帮助我。她不惜去除自己课间时间，来为我巩固课上认知识，带我做出家后整理一天中所错的题。后来我突然现，她体育不好，但这是我的强项啊！把中午所有的作业时间分割成她教我学习，监督她的体育，我们互帮互助，取长补短，不美哉！在我们的不懈努力之下，我们各自成绩都提高了不少。因为，在我们互相帮助时，即帮助了他人，又帮助了自己。希望我们的友谊能够天长地久。学校的走廊，永远是我们吹着晚风，谈论未来的小角落，长长的走廊，睹所有人的青春继续向前走吧，要相信彼方上有荣光在！

Prompt: In life, we meet all kinds of people: some command respect, others admiration, some touch our hearts, and some make us sigh. Write an essay titled "Such People Move Me," using at least 500 words. "Such people" may be someone you know, like a classmate or neighbor, or someone unfamiliar. It could be a specific person or a type of person **embodying a particular spirit or representing a trend**. Focus on their **individuality and qualities** to describe your **emotional inclination**.

Such People **Touch** Me

If friendship is an evergreen tree, then the sincere spring water from the heart must nourish it; if friendship is an everlasting flower, then it must be illuminated by the sun that rises in the heart. At the beginning of the first year, after several diagnostic tests, one girl stood out as special, consistently ranking at the top, respecting her teachers—a quiet and intelligent girl with both academic and moral excellence. This girl was moved by my outgoing personality and became good friends with me. My grades weren't great, and I often felt inferior, but I confided in her. Perhaps at that moment, she decided to help me. Sacrificing her break time, she helped me consolidate my knowledge and went through the questions I got wrong. Later, I realized that although she struggled with sports, it was my strength. So, we divided our lunch break into study time for her to tutor me and physical education time for me to coach her. Helping each other and learning from each other's strengths was a beautiful thing! With our unwavering efforts, both our grades improved significantly. When we helped each other, we not only assisted one another but also helped ourselves. I hope our friendship will last forever. The school corridor will forever be our little corner where we talk about the future in the evening breeze. As we all continue to move forward through our youth, let's believe there's glory ahead!

EduChat (a)

这篇作文的中心与切题方面表现良好。作者选取了一个具体的人物——尤唯，通过描写她的品质和行为，深刻地展现了她的精神风貌。尤唯是一个品学兼优、尊敬师长、文静聪慧的女孩，她的友谊和帮助让作者感动。作者通过描写自己与尤唯的互相互动，表达了对友谊的珍视和对未来的期许。整篇作文紧扣主题，情感真挚，给人以启发和感动。

GPT-4 (b)

作者选择了一位品学兼优、乐于助人的女孩作为写作对象，通过描写她的品质和行为，表达了对她的敬佩和感激之情。同时，文章也强调了友谊的重要性，以及互帮互助的精神。从切题的角度来看，这篇作文完全符合题目要求，以一个具体的人物为例，深入挖掘其精神品质和代表的风气，并表达了作者的情感倾向。因此，我认为这篇作文的中心与切题方面都做得很好。

Human (c)

这篇作文在中心和切题方面不合格。文章的题目是《这样的人让我感动》，应该侧重于写让作者感动的人的事迹，进而表现这个人的性格、品质等特征。本文却从友谊说起，讲二人之间是如何成为朋友，发展友情的，与题目要求不符。建议作者在写作前仔细审题，围绕中心组织事实材料。

The essay **excels** in focus and relevance. The writer depicts **Yu Wei**, a character chosen for her vivid spirit shown through qualities and actions. **Yu Wei** is presented as **academically and morally excellent, respectful, gentle, and intelligent**. Her friendship and assistance **touched** the writer, who reflects on **mutual support, valuing friendship, and hope for the future**. The piece stays on point with sincere emotions, inspiring and moving readers.

The author chose a girl who **excels academically and is eager to help others** as the subject for the essay, using descriptions of her qualities and actions to convey admiration and gratitude. The piece also highlights **the significance of friendship and the spirit of mutual assistance**. From the perspective of relevance, the essay **fully meets** the requirements, using a specific individual to delve into her moral fiber and the ethos she represents while **expressing the author's emotional stance**. Hence, I believe this essay **excels in both focus and relevance**.

The essay **fails** to meet the requirements in relevance and focus. The title is "Such People Touch Me" which should concentrate on the deeds of the person who moved the author, thereby showcasing their personality and qualities. **However, the essay starts with friendship, discussing how two people become friends and develop their relationship, deviating from the topic.** The author is advised **to carefully review the question before writing and organize facts around the central theme.**

Figure 5: An example of generating topic relevance comment for grade 7 student essay titled *Such People Move Me* using EduChat and GPT-4. *Human* denotes human-written comment for preference. Text highlighted in **bold** are points requiring special attention for students. The right side consists of comments generated through three models. In each side, the right part is the translated English version for better understanding. Texts highlighted in **red** indicate misinformation or incorrect advice, whereas texts highlighted in **blue** provide accurate feedback or suggestions for students. Specifically, in part (a), EduChat incorrectly assesses the level as "excel" and fabricates a character named Yu Wei that does not exist; in part (b), GPT-4 inaccurately evaluates the essay's level as "fully meets" and fails to identify the specific "author's emotional stance"; in part (c), human-written comment correctly points out the topic derivation of this essay, explain why, and provides detailed suggestions on how to refine it.

strong Chinese reading and writing skills. The selection process involved evaluating their test samples with BERTscore to determine their ability to understand and annotate text accurately.

Training and Division of Responsibilities   Before beginning the actual annotation work, the team underwent a training session to harmonize their understanding of the task's requirements and the annotation guidelines. This training ensured consistency across annotations. The team was then divided based on the educational level of the essays they would annotate: Three annotators focused on essays from primary school students (grades 3 to 6) and the remaining three handled essays from middle school students (grades 7 to 9). For each essay, one annotator was primarily responsible for the annotation, another served as a reviewer to ensure accuracy, and a third annotator was involved in case of significant discrepancies between the first two opinions.

Annotation Process Duration   The duration of the annotation process can vary widely depending on the volume of text to be annotated, the complexity of the task, and the team's efficiency. In our structured project, the entire process of annotation spanned a duration of one month.

Compensation   Compensation was provided to each annotator based on the actual volume of annotation work completed in a fair manner.

C.5   Explanation of Essay Genres and Prompts

As shown in Table 7, an explanation of 13 distinct literary genres are listed. Moreover, a complete list of **74** diverse essay prompts is shown in Table 9.

D   Prompt Template

Our prompt for two tasks is shown in Figure 7.

Level	Description
Excellent	The essay fully meets the requirements of the topic, the topic is very appropriate, the center is extremely clear, and the materials chosen are particularly suitable.
Good	The essay is somewhat in line with the requirements of the topic, the topic is somewhat appropriate, the center is relatively clear, and the materials chosen are relatively suitable.
Fair	The essay basically meets the requirements of the topic, the topic is generally appropriate, the center is somewhat clear, and the materials chosen are basically suitable.
Passable	The essay does not stray from the topic, the topic is slightly inappropriate, the center is not easy to extract, and the materials chosen are not very suitable.
Failed	The essay does not meet the requirements of the topic, the topic is very inappropriate, the center is extremely unclear, and the materials chosen are particularly unsuitable.

Table 6: Details of topic relevance level.

Figure 6: A screenshot of our annotation system.

E Full Results

Full results of automatic evaluation scores are displayed in Table 10.

F More Explanations on Error Analysis

As shown in Figure 4, we present the confusion matrices on Task 1.

- For GPT-4, its performance in predicting *excellent* is very good. However, for the remaining four levels, the predictions are generally overestimated, concentrated in the first column of the confusion matrix.

- For Baichuan2-7B, it can hardly make successful predictions, with a large number of predicted labels concentrated in the sixth column.

- For Baichuan2-7B-SFT, its prediction performance is better than Baichuan2-7B.

- For Baichuan2-7B-SFT-DPO, the prediction results are significantly more concentrated around the diagonal, and the over-prediction of higher-level labels is alleviated to some extent.

你是教学经验丰富的中学语文教师，请你为下面这篇[GRADE]年级学生的作文先点评切题等级，然后写中心与切题评语。请注意，等级要求为[LIST]，只点评文章的中心与切题方面，其他任何方面都不要点评。点评总字数不超过150字。作文题目要求为[PROMPT]，中心与切题点评要点为[REQUIREMENT]，作文题目为[TITLE]，作文内容为[CONTENT]。

You are an expert high school Chinese language teacher. Please first identify a level and then write a topic relevance comment for the following essay by [GRADE] grade students. Please note that details for levels are in [LIST] and only the central focus and on-topic aspects of the essay should be commented on. The total word count for the comment should not exceed 150 words. Essay Prompt: [PROMPT], Central Focus and topic relevance Comment requirements: [REQUIREMENT], Essay Title: [TITLE], Essay Content: [CONTENT]

Figure 7: The prompt used for generating topic relevance comment for an essay. The bottom part is the translated English version. Notably, [LIST] denotes the details of topic relevance levels shown in Table 6.

Genre	Description	Essay Prompt Title Example
Diary 日记	This genre focuses on personal entries arranged by date reporting on what has happened over the course of a day.	Rural life experience today 乡村生活记
Character 写人	The genre specializes in writing about people, often describing their characteristics, behaviours, experiences, personality, etc.	Such people touch me 这样的人让我感动
Fantasy 想象	This genre involves creating and writing about imaginary worlds, characters, and stories, often involving magic or other supernatural elements.	Me in ten years 十年后的我
Object 写物	A form of descriptive writing that focuses on objects, exploring their aspects in detail.	National treasure giant panda 国宝大熊猫
Scenery 写景	This genre is about describing the nature or environment. It captures the details and beauty of natural or urban landscapes in a literary manner.	Outside the window 窗外
Reflection 感想	This genre entails writing one's thoughts, feelings, and reflections on a certain topic, incident, or experience, especially after reading a book or watching a movie.	Reflections on a classic book 读后感
Narrative 记叙	This genre tells a story. It involves characters, a plot, and a storyline, and is focused on telling a continuous sequence of events.	An unforgettable event 难忘的一件事
Letter 书信	This genre is about writing personal or official letters in a literary manner, often used as a way of communication between individuals or organisations.	Mom, I want to tell you 妈妈，我想对您说
Proposal 倡议	This genre involves writing motivational or argumentative pieces that aim to encourage people to take up a specific belief, course of action, or idea.	Garbage sorting initiative 垃圾分类倡议书
Lyric 抒情	This genre specializes in expressing personal and emotional feelings, often written in a way intended to be spoken.	Nostalgia for hometown 乡情
Biography 传记	This genre is about writing about the lives and details of individuals, either an autobiography written by oneself or a biography written by another.	My father 我的父亲
Exposition 说明	This genre can involve explaining, defining, or interpreting a specific topic or idea.	Tablet computer 平板电脑
Argument 议论	This genre entails crafting a logical and structured argument on a specific topic, aiming to convince the reader of a particular viewpoint.	On contentment and happiness 论知足与快乐

Table 7: An explanation of 13 genres in TOREE. In the columns of *Genre* and *Essay Prompt Title Example*, the second line is the corresponding Chinese.

Data field	Content
instruction	There is a Chinese proverb, "Those who are close to vermillion are red, and those who are close to ink are black", which emphasizes the impact of environment on human growth.
	What do you think of it? Please make up your own mind, draw up your own topic, and write a refutation paper. No less than 600 words.
	Tips: 1. Sort out the way of writing the refutation, choose a suitable angle to refute, and at the same time show that your point of view is reasonable.
	2. Choose materials that are consistent with your point of view. It is best to have factual arguments and reasonable arguments.
	3. Make an outline based on your opinions and materials, communicate with classmates, supplement each other's arguments, and complete the composition on this basis.
essay	Title: Influence
	Content: There is a Chinese proverb, "Those who are close to vermillion are red, and those who are close to ink are black."
	What does this sentence mean? Some people say that this sentence is about the impact of environment on human growth.
	But I don't think so. My understanding is: if you play with good classmates or smart people, your grades will get better;
	if you play with bad classmates or people who are not smart, your grades will also get worse. That's right to understand!
	Growing up, parents would say: "Don't play with this person. He has poor grades. If you play with him, he will become worse and stupider.
	You should play with good classmates." When you think about it, What parents say is not unreasonable.
	I remember there was such a scene in a TV series: A female classmate who had very good academic performance and ranked first in the class met an unlearned and social
	friend outside. She knew how to drink and drink. smokes. ...Her friend will teach her things she didn't dare to do before.
	Gradually, they become the closest friends. This female classmate slowly learns to skip classes, drink, dress loosely, speak dirty words, and her grades will improve.
	It has plummeted, and the former number one fell to the bottom. Once, she was playing outside and met a male classmate she liked.
	Seeing her like this, the male classmate advised her to study hard and not to affect other classmates. Maybe this sentence irritated her, and she felt very unhappy.
	My beloved The man said something bad about her, and she felt resentful and took out a knife and pointed it at him. The male classmate didn't take it seriously and said she didn't dare.
	The female classmate stabbed the male classmate in a fit of rage.
	The male classmate went to the hospital, and the girl also went to jail. Looking at it this way, there are not such cases in reality. What parents say is not wrong. Friends and partners also have an impact on people's growth.
	Most people study or work in a day, and spend more than ten hours at school or in the company. In a week, you will work for three or four days, or even more. You will see classmates at school, and You will meet colleagues at the company, which is usually the case.
	The people you have the most contact with are not your parents but your classmates or colleagues. These are all friends or partners.
	Some of your behaviors may come from the influence of friends, and may even change dramatically. So can you say that friends and partners have no influence on a person's growth?
	That shouldn't be possible, the facts are before our eyes. Friends and partners do have an impact on people's growth.
topic_relevance_level	Failed
chosen_comment	This essay is unsatisfactory in terms of topic and focus.
	The requirement of the topic is to write a refutation paper, but the starting point of the author's refutation is not an essential objection to the sentence
	"Those who are close to red are red and those who are close to ink are black" , it just presents its meaning in another way, which does not meet the requirements of the topic.
	It is recommended that the author carefully review the topic and choose an angle that touches the foundation of this sentence to avoid going off topic.
rejected_comment	This essay is an argumentative essay. The theme is to refute the theme of "those who are close to red are red and those who are close to ink are black",
	that is, the impact of the environment on human growth, that is, a negative opinion on this view. The author believes that, If you play with good classmates or smart people,
	your grades will get better; if you play with bad classmates or people who are not smart, your grades will get worse.
	In the article, the author illustrates through examples that friends and partners are really important for people's growth.
	It has influence. However, the author did not fully prove his point of view. If he does not refute "Those who are close to red are red and those who are close to ink are black",
	it will be regarded as off topic.

Table 8: A data sample from TOREE test set.

Genre	# Essay Prompts	Essay Prompt Title
Diary	2	Rural life experience today; Write an observation diary
Character	10	Guess who he is; People with characteristics around us; My "self-portrait"; "Comic" teacher; All kinds of people; It's nice to have you; Such people _ me; My idol; _things; Show off the "talented people" in our class
Fantasy	12	I'll invent imagination; Wonderful imagination; Imagining like this is fun; Little "zoo"; A day with _; My fantastic ideas; A newly written story; My hometown in twenty years; The magical adventure; Transformation; Stories flowing from the pen tip; Me in ten years
Object	10	_makes life better; Write around a central idea; The colorful world in our eyes; My plant friends; National treasure giant panda; My animal friends; My beloved object; Introduce an object; Recommend a book; China's World Cultural Heritage
Scenery	7	This place is beautiful; Recommend a nice place; My paradise; scene of _; September; Outside the window; I love __ season
Reflection	2	Reflections on a classic book; Reflections after watching a film or drama
Narrative	18	That time was really fun; I did a little experiment; Kaleidoscope of life; Remembering a game; My heart pounding; Travel _; The moment I grew up; He _; Colorful activities; Growing up experiences; Debate; My troubles; Stories in the photo; My day; Festival; Travel notes; Beyond that mountain; My hometown
Letter	2	Write a letter; I want to tell you
Proposal	1	Learn to write a proposal
Lyric	1	Feelings for my hometown
Biography	1	A brief biography
Exposition	3	My life can't be without it; Introduce a technological product; The environment around me
Argument	5	Video games; Discuss on integrity; Contentment and happiness; The influence of the company one keeps; The colors of spring

Table 9: A list of **74** diverse essay prompt titles in TOREE. The symbol "_" signifies placeholders that are to be completed by students during the writing task. These blanks offer a structured yet flexible framework, encouraging creative expression as the students formulate their own content to fill in these gaps.

LLM	Task 1			Task 2				
	Precision	Recall	Weighted-F1	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L	BERTscore
Baichuan2-7B	**61.94**	13.89	20.31	4.79	17.81	3.62	13.89	60.24
w/ SFT	40.17	**50.00**	42.23	25.96	44.33	24.13	38.58	75.86
w/ SFT + DPO	<u>44.74</u>	<u>43.06</u>	<u>42.86</u>	<u>25.98</u>	<u>46.56</u>	<u>25.28</u>	**40.15**	<u>76.26</u>
Chinese-Alpaca2-7B	35.71	6.94	11.63	4.78	20.75	4.14	16.68	62.73
w/ SFT	39.14	40.28	36.64	24.27	43.91	22.84	37.72	74.80
w/ SFT + DPO	38.43	41.67	39.83	23.59	44.18	23.53	37.78	75.38
ChatGLM3-6B	29.41	13.89	18.87	4.78	17.49	4.35	12.31	59.27
w/ SFT	34.46	34.72	33.03	22.84	41.90	21.38	36.04	74.46
w/ SFT + DPO	34.72	19.96	33.49	23.96	44.02	22.28	36.27	75.05
Qwen-7B	37.74	27.78	30.82	10.45	31.46	8.59	23.27	69.27
w/ SFT	37.65	34.72	33.87	25.47	44.30	23.54	33.37	75.74
w/ SFT + DPO	40.79	**50.00**	**44.92**	**28.17**	**47.83**	**28.74**	<u>38.60</u>	**76.85**

Table 10: Full results of automatic evaluation scores on evaluated open-source LLMs. Within each part, best results are highlighted in **bold**, and second best results are <u>underlined</u>.